National Center for Biotechnology Information

https://www.ncbi.nlm.nih.gov/

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Creation of the National Center for Biotechnology Information (NCBI)

The late Senator Claude Pepper recognized the importance of computerized information processing methods for the conduct of biomedical research and sponsored legislation that established the National Center for Biotechnology Information (NCBI) on November 4, 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). The NLM was chosen for its experience in creating and maintaining biomedical databases and its ability to establish an intramural research program in computational molecular biology. Dr. David Lipman was selected to serve as the Director.

Who We Are and What We Do

The NCBI is made up of multidisciplinary research and development teams composed of molecular biologists, biochemists, computer scientists, mathematicians, research physicians, and structural biologists. The NCBI's mission is to perform basic and applied research and to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that affect health and disease. Since its inception, NCBI has become a leading international resource for public biomedical databases, software tools for analyzing molecular and genomic data, and research in computational biology.

Literature Databases

The NCBI's core literature database is **PubMed**, the most frequently consulted online scientific medical resource in the world. This service provides abstracts and citations for millions of articles from thousands of biomedical journals. In November of 2011 the PubMed database enabled the searching of over 21 million citations and abstracts from over 27,600 national and international medical and scientific journals. PubMed records link to full-text versions of the articles, when available, from journal websites or from **PubMed Central (PMC)**, an online archive of over 2.3 million free, full-text articles made available to the public in response to the NIH Public Access Policy. In addition to published manuscripts, PubMed records link to information from other databases including clinical literature, such as **OMIM** and **GeneReviews**, as well as the **Bookshelf**, a growing collection of biomedical books that can be searched electronically.

Contact: info@ncbi.nlm.nih.gov

Molecular Databases

The first national molecular database, **GenBank**, began modestly in 1982 as a collection of 600 DNA sequences funded by the NIH at the Los Alamos National Laboratory. In 1992, the coordination of nucleotide sequence submissions and hosting of the GenBank database was moved to the NCBI. The GenBank sequences are now incorporated into NCBI's **Nucleotide** database which includes all publicly available nucleotide sequences - currently over 212 million from more than 360,000 distinct organisms at or below the species level as well as a large number of uncultured organisms and organismal groups. Over the years, NCBI has created and managed more than 40 additional molecular databases housing biomolecular sequence and structural data, and the results of high-throughput experimental studies, such as those in the **SRA**, **dbSNP**, **GEO** and **PubChem** databases. In order to improve the use and impact of these large primary databases, NCBI has developed bioinformatics pipelines to produce curated resources. These include **RefSeq**, a curated reference collection of DNA, RNA, and protein sequences, **Gene**, a gene-specific database containing a mixture of curated and automated analyses integrating comprehensive information and links to related literature, molecular, genomic and biochemical data, and **BioSystems**, a biological pathways database providing links to associated literature, molecular, and biochemical data.

Computational Tools

Over the years, NCBI has created, maintained, and supported the use of over 40 databases for the medical and scientific communities as well as the general public. To enable users to find data in these databases, NCBI has developed the **Entrez search engine** to provide a unified and structured method for finding important data and also provides integrated access to the diverse set of biomedical literature, molecular and genomic information. In addition, the Entrez system was designed to provide links between related information across the NCBI databases to facilitate scientific and medical research and discovery efforts. In order to promote scientific discovery, new computational tools have been developed to assist in searching, analyzing and viewing data. These include the widely used **BLAST Sequence Similarity Search tool**, the **VAST** and **PubChem Structure Similarity Search** tools; data visualization applications, such as the **MapViewer**, a chromosome scaffolding, mapping and display tool, the **Sequence Viewer**, a graphical display window for sequences and annotations, **Genome Workbench**, a stand-alone application for viewing and analyzing user-derived and public sequence data; data clustering tools, such as **GEO Heatmaps**, a viewer for the examination of gene expression data, **CDTree**, a protein domain hierarchy viewer and editor; and the three dimensional molecular structure viewer and sequence alignment application **Cn3D**.

Recent Developments and Future Directions

NCBI is constantly improving and adding to its services as the needs of the scientific and medical community require. In late 2006, the **dbGaP** database, containing genotype data with phenotype information from NIH Genome-Wide Association Studies (GWAS), was launched and incorporated into the existing, set of literature and molecular databases. Recently, several new genomic structure and sequence databases were developed to support emerging scientific fields, including **dbVar**, a database of genomic structural variations, **RefSeqGene**, a collection of genomic sequences to be used as reference standards for mapping of annotations such as clinical sequence variations on well-characterized genes. **Genetic Testing Registry** (**GTR**) released to the public in early 2012 and catalogs information about genetic tests for inherited and somatic genetic variations, including arrays, multiplex panels, and pharmacogenetics. Newly released partner databases, **ClinVar** and **MedGen,** contain information regarding human variations genotypes and medically important phenotypes, and links to disease- and gene-specific information, and information related to medical genetics. The integration of all these sources of data continues the NCBI's tradition of facilitating identification of genetic factors that influence health, better understanding of disease, and support the development of effective clinical diagnostics and treatments. NCBI also collaborates with other government agencies, such as FDA and CDC, to provide services important to public health. **Pathogen Detection**, officially released in 2016, exemplifies the outcome of one of such efforts.

Public Outreach

NCBI conducts its public outreach through a variety of channels, such as webinars, workshops, social media, as well as talks, posters and exhibitions at scientific conferences. See the NCBI Learn page for more information https://www.ncbi.nlm.nih.gov/home/learn.shtml.